Classification of concepts described by taxonomic preordonnance variables with multiple choice

Israël-César Lerman, Philippe Peter

INRIA

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 1064

Programme 5
Automatique, Productique,
Traitement du Signal et des Données

CLASSIFICATION OF CONCEPTS DESCRIBED BY TAXONOMIC PREORDONNANCE VARIABLES WITH MULTIPLE CHOICE

Israël César LERMAN
Philippe PETER

Août 1989

RR_1064

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042-RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

CLASSIFICATION OF CONCEPTS DESCRIBED BY TAXONOMIC PREORDONNANCE VARIABLES WITH MULTIPLE CHOICE.

Israël-César LERMAN (IRISA-RENNES)

& Philippe PETER (IRESTE-NANTES)

Publication Interne n° 478

Juin 1989 - 16 Pages

ABSTRACT

Biological descriptions concerning phlebotomine sandflies of French Guiana are very complex (cf.[4]). Each species is a class of specimens and its description must represent not only a prototype but all the possible variations in the species. Thus, the description by a qualitative variable of a specimen representing a given species, requires -the most often- a disjunction of values. On the other hand, there are hierarchical relations on the set of variables. Finally, we suppose ranking similarity function on the modality set of each variable.

We formalize the expert knowledge base according to the previous description by means of a very new type of descriptor that we call : Taxonomic preordonnance variable with multiple choice.

A preordonnance variable (cf.[10]& [16]) is a qualitative one provided by a total preorder on the set of modality couples, interpreting the expert perceptual similarity between modalities. A taxonomic variable (cf.[2],[3], [10], & [15]) is a particular case of a preordonnance variable.

A taxonomic preordonnance with multiple choice variable (cf.[13] & [14]) is obtained by hierarchical organization of preordonnance variables where the "value" of a qualitative variable on a given object or on a given concept is a logical formula on the set of its modalities.

For such data, represented in a suitable way, we elaborate a similarity index between objects or concepts consistent with the Likelihood Linkage Analysis (L.L.A.) hierarchical classification method. This method enables the structuration of the family of species taken one by one or by couples (male, female).

In this context, the consensus problem (cf.[5]) appears as very particular.

This text is less technical and clearer than [14]. We distinguish here two notions. The first concerns the "value" of a variable on an object (e.g. specimen) and the second concerns the "value" of a variable on a concept (e.g. species). This classification takes into account discussion with J. Lebbe (cf.[4]).

Keywords : data in Artificial Intelligence, similarity index, likelihood linkage algorithm.

CLASSIFICATION DE CONCEPTS DECRITS PAR DES VARIABLES PREORDONNANCES TAXINOMIQUES A CHOIX MULTIPLE

RESUME

Les descriptions biologiques concernant les phlébotomes de la Guyane Française (cf.[4]) sont très complexes. Chaque espèce est une classe de spécimens et sa description doit représenter non seulement un prototype, mais toutes les variations possibles dans l'espèce. Ainsi, la description par une variable qualitative d'un spécimen représentant une espèce donnée, nécessite -le plus souvent- une disjonction de valeurs. D'autre part, il y a des relations de dépendance logique de nature hiérarchique sur l'ensemble des variables. Finalement, on suppose une relation de similarité ordinale sur l'ensemble des modalités d'une même variable.

Nous formalisons la base de connaissance de l'expert au moyen d'un nouveau type de descripteur que nous appelons : variable préordonnance taxinomique à choix multiple.

La variable préordonnance (cf.[10] & [16]) est une variable qualitative dont l'ensemble des couples de modalités est muni d'un préordre total qui exprime la perception des similarités entre modalités. Une variable taxinomique (cf.[2],[3],[10] & [15]) est un cas particulier de la variable préordonnance.

Une variable préordonnance taxinomique à choix multiple (cf.[13]& [14]) est obtenue à partir d'une organisation hiérarchique de variables préordonnance où la "valeur" de la variable qualitative sur un objet donné ou un concept donné est une formule logique sur l'ensemble de ses modalités.

Pour de telles données codées de façon adéquate, nous élaborons un indice de similarité entre objets (e.g. specimens) ainsi qu'entre concepts (e.g. species), conforme à la méthode de classification hiérarchique de l'"Analyse de la Vraisemblance du Lien" (A.V.L.). Cette méthode nous permet de structurer la famille des espèces.

Dans ce contexte, le problème du consensus en classification (cf.[5]) apparaît comme très particulier.

Ce texte reprend celui [14] en le dégageant de ses aspects les plus techniques et en le clarifiant, notamment en distinguant deux notions : "valeur" d'une variable sur un objet (e.g. specimen) et "valeur" d'une variable sur un concept (e.g. espèce). Cette clarification ne nous a été possible qu'après une discussion avec J. Lebbe (cf.[4]).

Mots-clés : Données en Intelligence Artificielle. Indice de similarité. Classification hiérarchique par l'Analyse de la Vraisemblance du Lien.

CLASSIFICATION OF CONCEPTS DESCRIBED BY TAXONOMIC PREORDONNANCE VARIABLES WITH MULTIPLE CHOICE.
APPLICATION TO THE STRUCTURATION OF A SPECIES SET OF PHLEBOTOMINE.

I.C. LERMAN (IRISA - RENNES)
& PH. PETER (IRESTE - NANTES)

INTRODUCTION

The "descriptive variable" notion can be formalized with much more generality and flexibility in Data Classification than in Factorial Analysis. In this latterfield a given variable has to be represented geometrically with respect to the whole set of the described objects. But in Data Classification the descriptive variable can have very synthetic expression of combinatorial and logical nature.

The underlied classification method is hierarchical and called "Likelihood Linkage Analysis". It has been set up and developed by I.C. Lerman & Collaborators (8). This method is based on a combinatorial and statistical approach and it leads to "significant" classes formation on both sets : the set of objects or concepts and the set of descriptive variables, and that, for all types of variables. Each variable is represented in the method, as a relation on the set of objects. We proceed to the interpretation of the results in a dual way by situating each class of objects with respect to the different variable classes and vice versa.

The purpose of this paper is to introduce a very new type of "descriptive variable" -or "descriptor" which is of combinatorial and logical nature and which has got a very synthetic and rich descriptive structure. This descriptor enables us to formalize the expert knowledge concerning the biological descriptions of phlébotomine sandflies of French Guiana [J. Lebbe, J.P. Dedet & R. Vignes (Institut Pasteur de la Guyane Française)(4)]. Descriptions are very complex. Each species is a class of specimens and its description must represent not only a prototype, but all possible variations in the species. Thus, the description by a qualitative variable of a given species, requires -the most often- a subset of possible values. Finally, we assume ranking similarity function on the modality set of each variable.

We will distinguish formally two description levels. The first concerns an elementary object x for which we only know that it belongs to a set $\mathcal{E}_C$ of "examples" of a concept C. The second level concerns the deduced description of the concept C. The problem is the to structure a set $\mathcal{P}$ of concepts. In our application x is a specimen and C is a species.

II. QUALITATIVE VARIABLE WITH UNIQUE CHOICE AND QUALITATIVE VARIABLE WITH MULTIPLE CHOICE

Let E be the set of objects or examples, where each object is a representative element of a concept C ; n=card(E). Let $J=\{1,2,\ldots,j,\ldots,m\}$ denote the coding of the modality set of a qualitative descriptive variable v. The structure of which the modality set is provided, is not taken into account in this paragraph. In classical data analysis v has to be defined on the whole set E ; but we consider here that v is a "descriptor" of E which an "unique choice", if there exists a non empty subset D of E, such that v is a mapping of D into J ; in other words $\mathcal{P}(E)$ denoting the set of subsets of E,

$$(\exists D \in [\mathcal{P}(E)-\{\emptyset\}]),\ (\forall x \in D) \Rightarrow v(x) \in J. \quad (1)$$

In fact by difference with the classical data classification, the "value" v(x) on the representation object x of a concept C, interprets the expert knowledge on C.

The descriptive variable v is "with multiple choice" on E, if there exists a non empty subset D of E, such that v is a mapping of D on a set of logical expressions on J. More precisely, each logical expression can be defined by a disjunction -eventually positively weighted by a probability distribution- of conjunctions, where each conjunction concerns a subset of J (i.e. a subset of modalities of J).

Thus, to a given object x of D, we can associate a sequence of subsets :

$$(J_1^x, J_2^x, \ldots, J_e^x, \ldots, J_k^x) \quad (2)$$

of the set J, provided by a probability distribution :

$$(p_1^x, p_2^x, \ldots, p_e^x, \ldots, p_k^x), \quad (3)$$

where the object x may possess the subset J_e^x of the modalities of the variable v, with the probability (or relative frequency) p_e^x $1 \leqslant e \leqslant k$. Notice that x may possess J_1^x or J_2^x or ... or J_k^x, where the "or" defines a strict disjunction. On the other hand, the different subsets J_e^x $(1 \leqslant e \leqslant k)$ are not necessarily disjoint.

We denote

$$(\exists D \in [\mathcal{P}(E) - \{\emptyset\}])(\forall x \in D),$$
$$v(x) = (\bigwedge\{j/j \in J_1^x\},\ p_1^x) \vee (\bigwedge\{j/j \in J_2^x\},\ p_2^x)$$
$$\ldots \vee (\bigwedge\{j/j \in J_k^x\},\ p_k^x). \quad (4)$$

We may now extend the variable or descriptor definition on the set $\mathcal{O}$ of the concepts C. Relative to (4) the "value" of v on the concept C that x represents is in fact a conjunction of the following form :

$$v(C) = (\bigwedge\{j/j \in J_1^C\}, p_1^C) \wedge (\bigwedge\{j/j \in J_2^C\}, p_2^C) \wedge$$
$$\ldots \wedge (\bigwedge\{j/j \in J_k^C\}, p_k^C), \quad (5)$$

where $(J_l^C, p^C) = (J_l^x, p^x)$ for $1 \leqslant l \leqslant k$, since the "value" J_1^C occurs with the relative frequency p_1^C and the "value" J_2^C occurs with the relative frequency p_2^C,...

III. TOTAL PREORDONNANCE STRUCTURE ON THE MODALITY SET OF A QUALITATIVE VARIABLE. REPRESENTATION

A "preordonnance" qualitative variable is a qualitative variable of which the modality set J is provided by a total preordonnance which is a total preorder on the set of unordered (or ordered) modality couples. In the case of interest it concerns here the following set :

$$J^{\{2\}} = \{(j,h)/1 \leqslant j \leqslant h \leqslant m\} \quad (1)$$

(cf. [10] & [16]). This total preorder is established directly by the expert or by means of an Artificial Intelligence program taking into account the expert knowledge and a theory of the concerned field. The total preorder is esta-

blished from the highest couples to the lowest ones, according to an ordinal comparison of the similarities on $J^{\{2\}}$ {i.e.$(j,h)\succ(j',h')$ [resp.$(j,h)\sim(j',h')$] if j and h are strictly more [resp. equally similar than j' and h'].

This last total preorder is coded by means of a "ranking function" which leads to the table

$$\{r_{jh}/(j,h)\in J^{\{2\}}\}, \quad (2)$$

where the rank r_{jh} is computed with the following formula :

$$r_{jh}=l_1+l_2+\ldots+l_{(p-1)}+\frac{1}{2}(l_p+1), \quad (3)$$

where l_q denotes the q^{th} class cardinal of the total preorder on $J^{\{2\}}$ and where (j,h) belongs to the p^{th} class, according to an increasing ordering of the preorder classes.

Example. Let consider the 33rd variable of phlebotomine sandflies description (cf.[4]), "Aspect of individual duct". The modalities of this variable are :

1- Smooth non-sclerotized
2- Smooth sclerotized
3- Transversely striated or annulated
4- With small prominent tubercles.

By going from the most similar couples to those, the least similar, the expert has given the following preordonnance on J={1,2,3,4} :

$$11\sim22\sim33\sim44\succ12\sim13\sim23\succ14\sim24\sim34, \quad (4)$$

where jh represents the pair {j,h}, $1\leq j\leq h\leq 4$.

Then -by ordering lexicographically the couples (j,h)- the table (2) becomes :

$$\{8.5,5,5,2,8.5,5,2,8.5,2,8.5\} \quad (5)$$

IV. TAXONOMIC VARIABLE ORGANIZING A SET OF QUALITATIVE VARIABLES. REPRESENTATION

Such variable that we denote ω consists of a sequence of sequences of qualitative variables of the following form :

$$\omega=\{v^{(1)};v_1^{(2)},\ldots,v_{k_2}^{(2)};\ldots;v_1^{(p)},\ldots,v_{k_p}^{(p)};\ldots;v_1^{(q)},\ldots,v_{kq}^{(q)}\}. \quad (1)$$

The first sequence is necessarily reduced to one element : the variable $v^{(1)}$ which is associated to the root of the tree figuring the taxonomic variable. On the one hand, each variable $v_{j_p}^{(p)}$ $(1\leq j_p\leq k_p)$ is the "daughter" variable of at least one variable $v_h^{(p-1)}(1\leq h\leq k_{p-1})$ and on the other hand, when two variables $v_j^{(p)}$ and $v_{j'}^{(p)}$ have the same "mother" variable $v_h^{(p-1)}$, they are respectively defined on two distinct subsets where each subset is characterized by one modality of $v_h^{(p-1)}$.

Example. Let consider the variables 1,18,19 and 20 of [4] ; where v^1 is the "Sex", v^2 is the "Number of style spines", v^3 is the "Distribution of insertion of 4 style spines" and v^4 is the "Distribution of insertion of 5 style spines".

We obtain the following taxonomic structure :

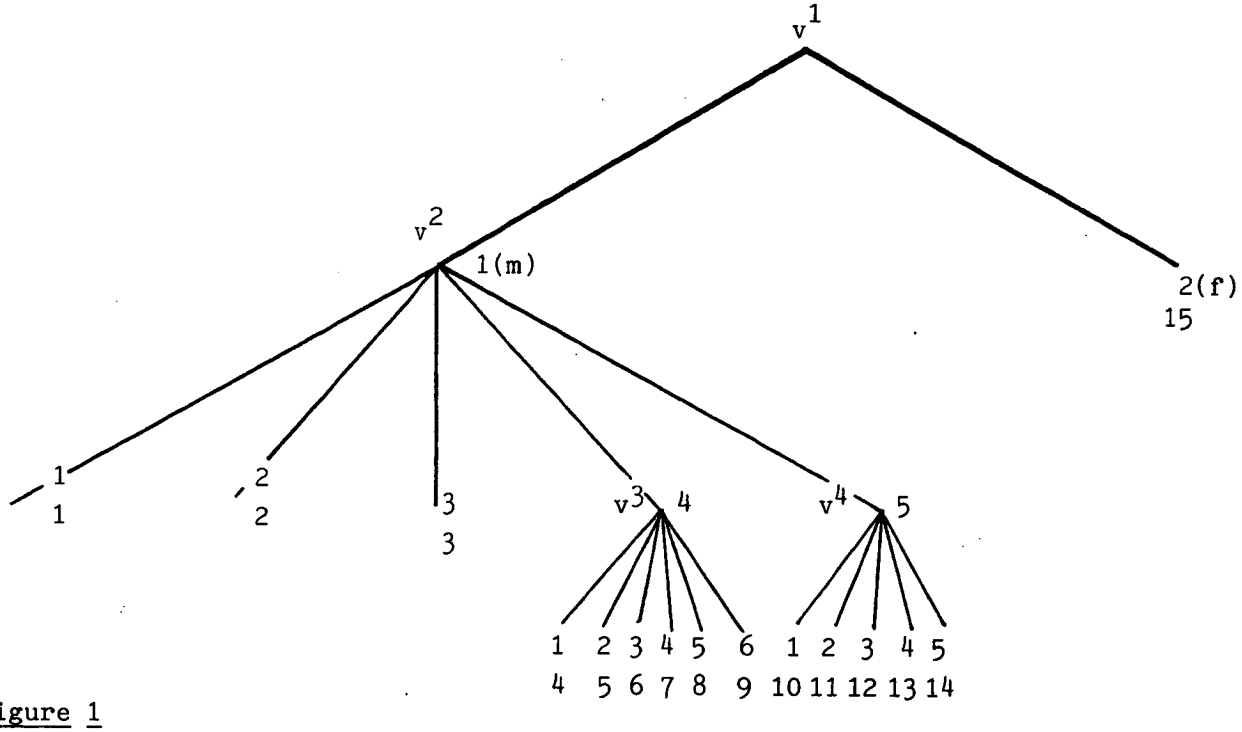

Figure 1

We represent the variable defined by this structure by means of what we call an "ultrametric preordonnance" (cf.[7] & [10]) on the set of the taxonomy leaves (the cardinality of this set is 15 in the above example). According to this total preorder on the set of unordered leaf couples, the higher is the rank of a given couple, the lower is the first node -for a decreasing construction of the taxonomic tree from the root to the leaves- which underlies the two leaves. Thus, in the above example, the pair {6,8} has the same rank as the pair {10,12}. This last is greater than that one of { 7,12}, which is equal to the {2,3} rank, and so on...

Let L denote the set of the leaves, a ranking function r coding the ultrametric preordonnance is characterized by the following property :

$$(\forall \{x,y,z\} \in P_3(L)),\ r(x,z) \geq \min[r(x,y),(r(y,z)]$$

We adopt as in the general case (cf.§III), the notion of "mean rank" which can be determined by a formula from the "type" of the taxonomy (cf. [14]). We must precise that the preordonnance class at which is assigned the highest rank is composed of the couples (x,x), $x \in L$.

By this way, the taxonomic variable appears as a particular case of the preordonnance variable.

V. TAXONOMIC PREORDONNANCE VARIABLE. REPRESENTATION

Relative to the above defined structure of a taxonomic variable ω, we further assume that the set $M_{j_p}^{(p)}$ of the modalities of a qualitative variable $v_{j_p}^{(p)}$, is provided by a total preordonnance expressing -in a ordinal way- the modality resemblances ($1 \leq j_p \leq k_p$, $1 \leq p \leq q$). On the other hand, we suppose that these k_p preordonnances can be extended in an unique total preorder on the set :

$$\bigcup \{P_2(M_{j_p}^{(p)}) / 1 \leq j_p \leq k_p\}, \quad (1)$$

according to an ordinal similarity on the modality pairs, where $P_2(X)$ denotes the unordered object pairs of X.

In these conditions, we have to define a total preordonnance on the set of the leaves of the taxonomy, or -equivalently- on the set of complete chains, going from the root to the leaves. This preordonnance must take into account both the preordonnance defined in the previous paragraph IV, and those that we have just introduced.

Such preordonnance is built step by step, in a decreasing way with respect to -perceived or deduced- resemblance between terminal modalities represented by the leaves of the taxonomy. Then, the first class of the preordonnance concerns the pairs {x,x}, where x is a leaf modality ; x belongs to {1,2,...,15} in the above paragraph IV example. The general principle consists of refining the ultrametric preordonnance associated to the taxonomy by means of the preordonnances, locally defined. Each one of these last concerns the set of the node pairs separated at a given taxonomy level and joined at the consecutive level. Thus, in the above example, if A={4,5,6,7,8,9} and B={10,11,12,13,14}, the continuation of the preordonnance will concern $[P_2(A) \cup P_2(B)]$, where $P_2(A)$ [resp.$P_2(B)$] denotes the unordered element pairs of A (resp. B). Now, if C denotes the set of the leaves {1,2,...,14}, the continuation of the preordonnance will concern $\{P_2(C)-[P_2(A) \cup P_2(B)]\}$.

The finally obtained structure is a total preordonnance on the set of the taxonomy leaves, that we code by means of a ranking function [see formula (3) §III].

VI. PREORDONNANCE TAXONOMIC VARIABLE WITH MULTIPLE CHOICE

The description structure of the variable that we consider here is identical to that one defined in the preceding paragraph V ; but now, the "value" of a given variable $v_{j_p}^{(p)}$ ($1 \leq p \leq q, 1 \leq j_p \leq k_p$) on a given concept is defined by a weighted disjunction of conjunctions on the set $J = M_{j_p}^{(p)}$ of its modalities ; that is to say, an expression as that (4) of paragraph II above.

By referring to this logical expression, description [1] can be formalized by the particular case where J_e^x includes only one element, and where $p_e^x = 1/k$, $1 \leq e \leq k$. Besides, description [4] can be expressed by the case where the sequence of subsets $(J_1^x, J_2^x, \ldots, J_k^x)$ is reduced to one non empty subset J^x, strictly included in J.

The "value" of such variable on an object representing a given concept will finally correspond to a weighted disjunction of conjunctions. Each conjunction concerns a subset of terminal leaves, or -equivalently- of complete chains, joining the root to the terminal leaves. The weight assigned to a conjunction of complete chains will be computed multiplicatively, with respect to their respective compositions.

Example : Let us imagine that -for a given concept x- we have for the descriptive variables introduced in the paragraph IV example :

$$v_2^1(x)=1$$
$$v_2^2(x)=(1\&2,0.4)\vee(2\&3\&4,0.2)\vee(4\&5,0.4)$$
$$v_3^3(x)=(1\&2,0.4)\vee(2\&3,0.6)$$
$$v^4(x)=(2\&3\&4,0.8)\vee(3\&5,0.2),$$

where & denotes a conjunction.

A given path is represented by the sequence of the modalities of the different variables organized according to the taxonomy. The value of the "taxonomic preordonnance variable with multiple choice" ω, is :

$$\omega(x)=(11\&12,0.4)\vee(12\&13\&141\&142,0.08)\vee$$
$$(12\&13\&142\&143,0.12)\vee(141\&142\&152\&153\&154,0.128)\vee$$
$$(141\&142\&153\&155,0.032)\vee(142\&143\&152\&153\&154,0.192)\vee$$
$$(142\&143\&153\&155,0.048);$$

or, by using the coding of the taxonomy leaves with the integers 1 to 15 (cf. Figure 1),

$$\omega(x)=(1\&2,0.4)\vee(2\&3\&4\&5,0.08)\vee(2\&3\&5\&6,0.12)\vee$$
$$(4\&5\&11\&12\&13,0.128)\vee(4\&5\&12\&14,0.032)\vee$$
$$(5\&6\&11\&12\&13,0.192)\vee(5\&6\&12\&14,0.048)$$

We can check that the weighting sum is equal to unity. This point -as we may show- is general.

Now, the "value" of v on the concept C represented by x, can be put in the following form :

$$\omega(C)=(1\&2,0.4)\wedge(2\&3\&4\&5,0.08)\wedge(2\&3\&5\&6,0.12)\wedge$$
$$(4\&5\&11\&12\&13,0.128)\wedge(4\&5\&12\&14,0.032)\wedge$$
$$(5\&6\&11\&12\&13,0.192)\wedge(5\&6\&12\&14,0.048)$$

where $\wedge$ is another notation for a conjunction.

VII. ROUGH SIMILARITY INDEX BETWEEN TWO OBJECTS (resp. CONCEPTS) DESCRIBED BY A TAXONOMIC PREORDONNANCE VARIABLE WITH MULTIPLE CHOICE.

Let us recall the case of a preordonnance variable with unique choice (cf. §III,[10] & [16]). To classify the object set, we have to consider ordinal similarity as symmetrical notion coded by the ranking function introduced in paragraph III cf. (2) & (3) §III).

Let x and y be two given objects. If ω denotes the preordonnance variable with unique choice and if $\omega(x)$ [resp. $\omega(y)$] is the value of ω on x [resp. y] [i.e. the modality of ω possessed by x (resp. y)], then the rough similarity index between x and y, is

$$s(x,y)=r[\omega(x),\omega(y)] \quad (1)$$

Now, let us consider the case where ω is a preordonnance variable with multiple choice. The value of the variable ω on the object x (resp. y) has the form (4) of the above paragraph II.

According to the notations of this last paragraph, let us denote by

$$\{(J_e^x,p_e^x)/1\leqslant e\leqslant k\} \quad (2)$$

the components of the weighted disjunction corresponding to $\omega(x)$.

On the other hand

$$\{(J_f^y,p_f^y)1\leqslant f\leqslant l\} \quad (3)$$

are the components of the weighted disjunction corresponding to $\omega(y)$.

The estimation of the similarity between x and y requires matching coefficient between (J_e^x,p_e^x) and (J_f^y,p_f^y), $1\leqslant e\leqslant k$, $1\leqslant f\leqslant l$.

Let us introduce, for comparing J_e^x and J_f^y :

$$(\forall j \in J_e^x), r(j)=\max\{r(j,h')/h' \in J_f^y\} \quad (4)$$

$$(\forall h \in J_f^y), r(h)=\max\{r(j',h)/j' \in J_e^x\} \quad (5)$$

Then, the contribution that we consider of (J_e^x,J_f^y) to the rough similarity index between x and y, is

$$s_{ef}(x,y)=\mathcal{M}\{\{r(j)/j\in J_e^x\},\{r(h)/h \in J_f^y\}\}, \quad (6)$$

where $\mathcal{M}$ represents mean operation over $[\mathrm{card}(J_e^x) + \mathrm{card}(J_f^y)]$ integers.

The integration of the whole following set of weighted couples of J subsets [cf.(7) below] -to evaluate global rough index s(x,y)- is done by a mean operation $\mathcal{M}^p$ [cf. (8) below], with respect to the distribution (9) :

$$\{[(J_e^x,p_e^x),(J_f^y,p_f^y)\ /1\leqslant e\leqslant k,\ 1\leqslant f\leqslant l\}, \quad (7)$$

$$s(x,y)= \mathcal{M}^p\{s_{ef}(x,y)/1\leqslant e\leqslant k,\ 1\leqslant f\leqslant l\}, \quad (8)$$

$$\{p_e^x \times p_f^y/1\leqslant e\leqslant k,\ 1\leqslant f\leqslant l\} \quad . \quad (9)$$

Let us now define the rough similarity index $s(C^x,C^y)$ between the two concepts C^x and C^y that x and y represent respectively. $\omega(C^x)$ [resp. $\omega(C^y)$] can be expressed by (5) of paragraph II, where $\{(J_d^C,p_d^C)/1\leqslant d\leqslant k\}$ is replaced by $\{(J_e^{C^x},p_e^{C^x})/1\leqslant e\leqslant k\}$ [resp. $\{(J_f^{C^y},p_f^{C^y})/1\leqslant f\leqslant l\}$]

Taking into account the above expression (6) -which matches J_e^x and since $\omega(C^x)$ [resp. $\omega(C^y)$] is defined by a conjunction, we have to determine for each e ($1 \leqslant e \leqslant k$) [resp. f ($1 \leqslant f \leqslant l$) the best matching of J_e^{Cx} with a J_f^{Cy} (resp. of J_f^{Cy} with a J_e^{Cx}). Namely, we define

$$s_{ef(e)}(x,y)=\max\{s_{ef'}(x,y)/1 \leqslant f' \leqslant l\}, \qquad (10)$$

and

$$s_{e(f)f}(x,y)=\max\{s_{e'f}(x,y)/1 \leqslant e' \leqslant k\} . \qquad (11)$$

The rough similarity index that we adopt between the two concepts C^x and C^y can be written as follow :

$$s(C^x,C^y)=\mathcal{M}_e^p\{\{s_{ef(e)}(x,y)/1 \leqslant e \leqslant k\}, \{s_{e(f)f}(x,y)/1 \leqslant f \leqslant \}\} \qquad (12)$$

where $\mathcal{M}_e^p$ denotes mean operation with respect to the following distribution

$$\{\{p_e^x p_{f(e)}^y/1 \leqslant e \leqslant k\}, \{p_{e(f)}^x p_f^y/1 \leqslant f \leqslant l\}\}. \qquad (13)$$

VIII. SIMILARITY STRUCTURE BETWEEN OBJECTS (CONCEPTS) COMPATIBLE WITH LIKELIHOOD LINKAGE CRITERION IN CASE OF SEVERAL VARIABLES.

We have studied this similarity structure in [10]. Let E denote here a set of objects or a set of concepts and let $\mathcal{W}$ be a set of descriptive variables of any type. Relative to a given variable $\omega(\omega \in \mathcal{W})$, we have built the rough index $s_\omega(x,y)$ between two elements x and y of E, in case where ω is a preordonnance variable with multiple choice and in case where E is a set of objects [cf. (8) above] or a set of concepts [cf. (12) above].

In our approach we normalize the contribution $s_\omega(x,y)$ with respect to the set of unordered element pairs of E : $F=P_2(E)$. This normalization is statistical ; namely, we define :

$$S_\omega(x,y)= \frac{[s_\omega(x,y)-moy_e(s)]}{\sqrt{var_e(s)}} \qquad (1)$$

where $moy_e(s)$ and $var_e(s)$ are respectively the empirical mean and variance of $s_\omega(x,y)$ over $F=P_2(E)$.

To evaluate the similarity relative to the whole set $\mathcal{W}$ of the variables, we begin by considering

$$S(x,y)= \frac{1}{\sqrt{card(\mathcal{W})}} \sum \{S_\omega(x,y)/\omega \in \mathcal{W}\} \qquad (2)$$

which is statistically standardized over $F=P_2(E)$. We obtain :

$$S_\nu(x,y)= \frac{[S(x,y)-moy_e(S)]}{\sqrt{var_e(S)}}, \qquad (3)$$

where $moy_e(S)$ and $var_e(S)$ are respectively the empirical mean and variance of $S(x,y)$ over $F=P_2(E)$.

At this level, we consider a hypothesis of independence or no relation for which we associate to the set of observed variables $\mathcal{V}$, a set $\mathcal{V}^*$ of independent random variables, respectively of the same types as those of $\mathcal{V}$. In these conditions the computed similarity index (3), $S^*_v(x,y)$, based on $\mathcal{V}^*$, is a random variable. In case where the number of variables, $card(\mathcal{V})$ is not too small, it can be shown that $S^*_v(x,y)$ is normally $N(0,1)$ distributed. Then the similarity coefficient that we adopt finally takes the following form

$$P(x,y)= \Phi[S_v(x,y)] \qquad (4)$$

where Φ is the cumulativedistribution function of the standardized normal variable.

This similarity index is referring to a probability scale. The similarity table is then

$$\{P(x,y)/\{x,y\} \in F=P_2(E)\}. \qquad (5)$$

The agglomerative building of hierarchical classification tree on E, is based in our method on the "likelihood of the maximal linkage" criterion that we have introduced in 1970 [cf. (6),(8)]. This criterion has got clear statistical basis with respect to statistical independence hypothesis between the two disjoint subsets C and D (defining two classes) of E, to be compared at a given level of the tree construction. Its expression is the following :

$$P(C,D)=[\max\{P(x,y)/(x,y)\in CxD\}]^{cxd} \qquad (6)$$

where c=card(C) and d=card(D).

Computational reasons lead us to use the strictly increasing function on the interval [0,1] :

$$f(\xi)=-Log[-Log(\xi)] \qquad (7)$$

and technically to adopt the equivalent criterion

$$Q(C,D)=-Log\{-Log[P(C,D)]\} \qquad (8)$$

In order to build the classification tree, we have recently established reactualization formulas in case of multiple aggregations at a given level for the most classical hierarchical classification criteria ("Single Linkage", "Complete Linkage", "Average Linkage", "Ward Criterion" and "Likelihood Maximal Linkage Criterion") (cf.[11]).

We note in passing that the "Likelihood Maximal Linkage Criterion" has got the "Reductibility" property [cf.(8)] introduced by Bruynooghe in 1977. This property makes possible algorithm implementation for classification of large data set [cf.(1)].

A decisive stage of the "Likelihood Linkage Analysis" (L.L.A.) classification method consists of the recognition of the "Significant" levels and nodes of hierarchical classification tree, that we reduce to the levels where "significant" nodes appear [cf.(7),(8) & (9)]. This is done by means of an association criterion between the emerged partition at a given level anda suitable structure of the information concerning the resemblances between the elements of the set E to be classified. This criterion that we have assessed from both points of view, formal and statistical, can be based on the preordonnance on E associated to the Similarity function (on E) or directly on the Similarity function itself [cf.(8) & (9)]. The recognition of the most "significant" levels and the most "significant" nodes is based on the behaviour of the distribution of the mentionned criterion, on the increasing sequence of the levels of the classification tree. A "significant" level indicates natural partition to consider with a given degree of synthesis. The significant nodes indicate completion stages of the different classes appearing in the tree.

We show in the following figure 2 a branch of the hierarchical classification tree that we have obtained on a set of phlebotomine sandflies species that we have described by means of taxonomic preordonnance variables with multiple choice. In this structure, the significant nodes are represented by the darker and thicker lines.

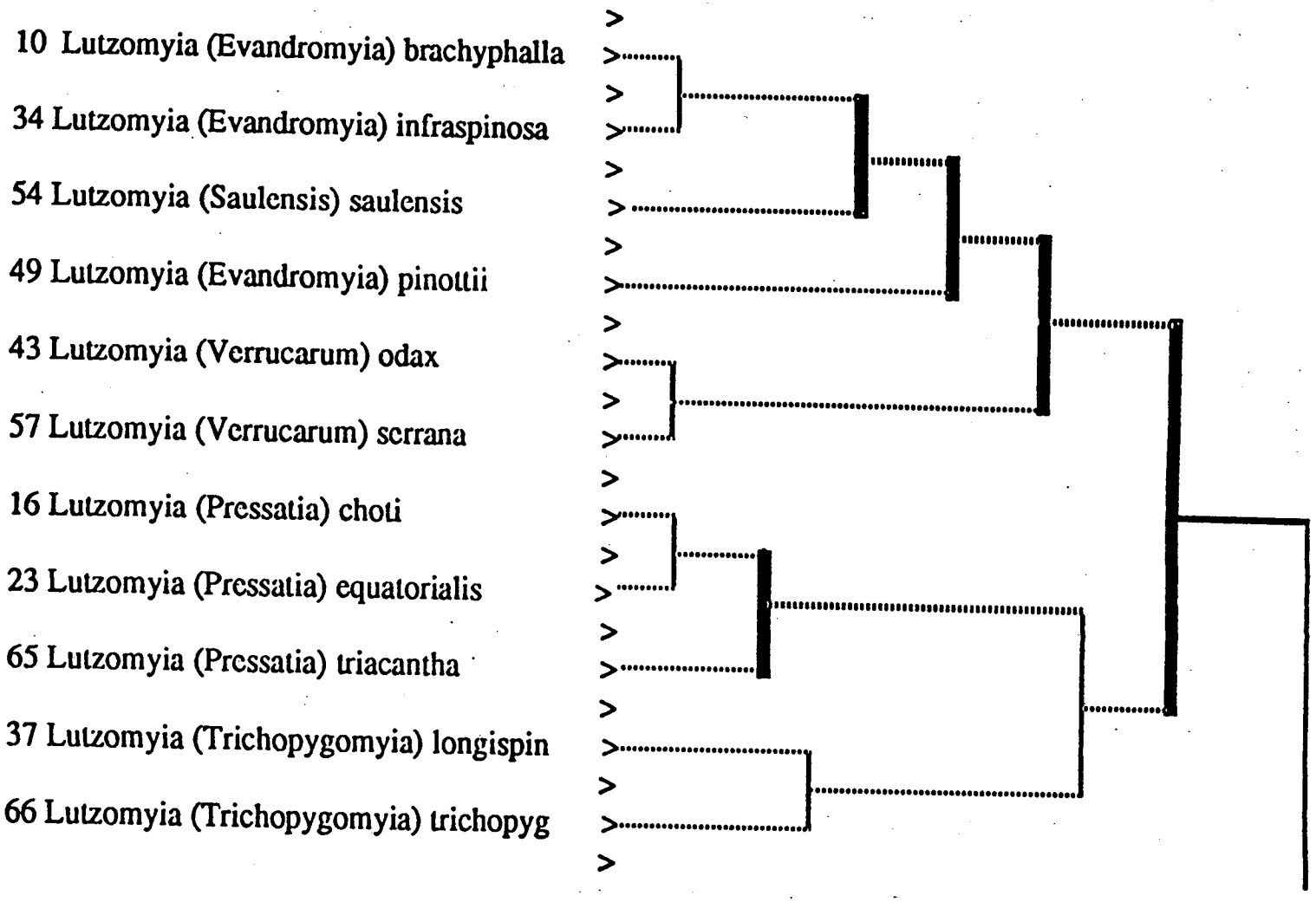
10 Lutzomyia (Evandromyia) brachyphalla
34 Lutzomyia (Evandromyia) infraspinosa
54 Lutzomyia (Saulensis) saulensis
49 Lutzomyia (Evandromyia) pinottii
43 Lutzomyia (Verrucarum) odax
57 Lutzomyia (Verrucarum) serrana
16 Lutzomyia (Pressatia) choti
23 Lutzomyia (Pressatia) equatorialis
65 Lutzomyia (Pressatia) triacantha
37 Lutzomyia (Trichopygomyia) longispin
66 Lutzomyia (Trichopygomyia) trichopyg

Figure 2

REFERENCES

[1] BRUYNOOGHE M. (1989) ; "Nouveaux algorithmes en classification automatique applicables aux très grands ensembles de données, rencontrés en traitement d'images et en reconnaissance des formes", Thèse de Doctorat d'Etat, Univ. de Paris VI, 23 Janv. 1989.

[2] DECAESTECKER C. (1988) ; "Conceptual clustering with hierarchical and conjunctive variables", Rapport technique I.A. -03-88, CADEPS, Univ. Libre de Bruxelles.

[3] DIDAY E. (1987) ; "Introduction à l'approche symbolique en analyse des données", Journées "Symbolique-Numérique", Université Paris 9 Dauphine, 8-9 déc. 87.

[4] LEBBE J., DEDET J.P. & VIGNES R. (1987) ; "Identification assistée par ordinateur des phlébotomes de la Guyane Française", Institut Pasteur de la Guyane Française, Version 1.02 - 10/07/1987.

[5] LECLERC B. & CUCUMEL G. (1987) ; "Consensus en Classification : une revue bibliographique", Rev. Math. & Sc. Humaines 100, p. 109-118.

[6] LERMAN I.C. (1970a) ; "Sur l'analyse des données préalable à une classification automatique ; proposition d'une nouvelle mesure de similarité", Rev. Math. & Sc. Humaines n° 32, p. 5-15.

[7] LERMAN I.C. (1970b) ; "Les bases de la classification automatique", Gauthier Villars "collection Programmation", Paris.

[8] LERMAN I.C. (1981) ; "Classification et analyse ordinale des données", Dunod, Paris.

[9] LERMAN I.C. (1983) ; "Sur la signification des classes issues d'une classification automatique" in Numerical Taxonomy, NATO ASI series vol. G1, edited by J. Felsenstein, Springer Verlag (1983), p. 179-198.

[10] LERMAN I.C. (1987) ; "Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification". Rev. de Stat. Appl., XXXV (2), 39-60.

[11] LERMAN I.C. (1988) ; "Formules de réactualisation en cas d'agrégations multiples", Rapport Interne IRISA n° 409. A paraître dans RAIRO-R.O. (Vol. 23, n° 2, 1989).

[12] LERMAN I.C., HARDOUIN M. et CHANTREL T. (1980) ; "Analyse de la situation relative entre deux classifications floues", in Data Analysis and Informatics, E. Diday et al. (eds), North Holland (1980).

[13] LERMAN I.C., NICOLAS J. & PETER Ph. (1988) ; "Classification Conceptuelle: une approche dans la combinaison de méthodes numériques et symboliques", Deuxièmes Journées "Symbolique-Numérique" pour l'Apprentissage de Connaissances à partir de Données, Orsay, 12-13 Décembre 1988.

[14] LERMAN I.C. & PETER Ph. (1988) ; "Classification en présence de variables préordonnances taxonomiques à choix multiple. Application à la structuration des phlébotomes de la Guyane Française", Public. Int. n° 426 IRISA, Rennes, Septembre 1988.

[15] MICHALSKI R.S. (1980) ; "Pattern recognition as rule-guided inductive inference", IEEE Transactions on pattern analysis and machine intelligence, vol. PAMI-2, n° 4, July 1980.

[16] PETER Ph. (1987) ; "Méthodes de classification hiérarchiques et problèmes de structuration et de recherche d'informations assistées par ordinateur", Thèse de l'Univ. de Rennes I, 6 mars 1987.

LISTE DES DERNIERES PUBLICATIONS INTERNES IRISA

PI 472

PI 472 **DERIVATION SYSTEMATIQUE D'UN ALGORITHME DE SEGMENTATION D'IMAGES - UN EXEMPLE D'APPLICATION DU FORMALISME GAMMA**
Christian CREVEUIL, Gersan MOGUEROU
46 Pages, Mai 1989.

PI 473 **MICROCODE OPTIMIZATION FOR THE PCS PROCESSOR**
François BODIN, François CHAROT, Charles WAGNER
26 Pages, Mai 1989.

PI 474 **ALGEBRAICALLY CLOSED THEORIES**
Eric BADOUEL
22 Pages, Mai 1989.

PI 475 **QUELQUES OUTILS GRAPHIQUES POUR LA MODELISATION DU CONTROLE D'EXECUTION EN ROBOTIQUE DE COOPERATION**
Jean-Christophe PAOLETTI, Lionel MARCE
52 Pages, Juin 1989.

PI 476 **SIMULATION REPARTIE DE SYSTEMES A EVENEMENTS DISCRETS: PARTIE 1 : MODELISATION ET SCHEMAS D'EXECUTION**
Philippe INGELS, Michel RAYNAL
26 Pages, Juin 1989.

477 **PROGRAMMING WITH MALI - UNIFICATION OR ORDERED TYPES**
Olivier RIDOUX
18 Pages, Juin 1989.

PI 478 **CLASSIFICATION OF CONCEPTS DESCRIBED BY TAXONOMIC PREORDONNANCE VARIABLES WITH MULTIPLE CHOICE**
IsraëlCésar LERMAN
16 Pages, Juin 1989.

PI 479 **A SIMPLE GRAPH CONSTRUCTION OF SEMILINEAR REACHABILITY SETS OF VECTOR ADDITION SYSTEMS**
Gilles LESVENTES
16 Pages, Juin 1989.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique